# The cognitive plausibility of statistical classification models

Klavan, Jane; Divjak, Dagmar

*DOI:*
10.1515/flin-2016-0014
https://www.degruyter.com/view/journals/flin/50/2/article-p355.xml



*Document Version*
Publisher's PDF, also known as Version of record

*Citation for published version (Harvard):*
Klavan, J & Divjak, D 2016, 'The cognitive plausibility of statistical classification models: comparing textual and behavioral evidence', *Folia Linguistica*, vol. 50, no. 2, pp. 355-384. https://doi.org/10.1515/flin-2016-0014, https://doi.org/https://www.degruyter.com/view/journals/flin/50/2/article-p355.xml

Survey article

Jane Klavan* and Dagmar Divjak

# The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence

DOI 10.1515/flin-2016-0014
Submitted June 1, 2015; Revision invited November 17 2015;
Revision received February 29, 2016; Accepted May 31, 2016

**Abstract:** Usage-based linguistics abounds with studies that use statistical classification models to analyze either textual corpus data or behavioral experimental data. Yet, before we can draw conclusions from statistical models of empirical data that we can feed back into cognitive linguistic theory, we need to assess whether the text-based models are cognitively plausible and whether the behavior-based models are linguistically accurate. In this paper, we review four case studies that evaluate statistical classification models of richly annotated linguistic data by explicitly comparing the performance of a corpus-based model to the behavior of native speakers. The data come from four different languages (Arabic, English, Estonian, and Russian) and pertain to both lexical as well as syntactic near-synonymy. We show that behavioral evidence is needed in order to fine tune and improve statistical models built on data from a corpus. We argue that methodological pluralism is the key for a cognitively realistic linguistic theory.

**Keywords:** statistical modeling, near-synonymy, corpus linguistics, (psycho)linguistic experiments

## 1 Introduction

The use of probabilistic statistical classification models in linguistics was pioneered by Sankoff and Labov (1979). Fitting models to predict constructional and lexical choice is a growing trend in usage-based linguistics. It is a

***Corresponding author: Jane Klavan,** Department of English Studies, University of Tartu, Ülikooli 18, 50090 Tartu, Estonia, E-mail: jane.klavan@gmail.com
**Dagmar Divjak,** School of Languages & Cultures, University of Sheffield, Jessop West, 1 Upper Hanover Street, Sheffield S3 7RA, UK, E-mail: d.divjak@sheffield.ac.uk

method widely applied in semantics (e. g., Arppe and Järvikivi 2007; Arppe 2008; Divjak 2010; Divjak and Arppe 2013), syntax (e. g., Gries 2003; Bresnan 2007; Bresnan et al. 2007; Bresnan and Ford 2010; Kendall et al. 2011; Klavan 2012), morphology (e. g., Antić 2012; Baayen et al. 2013), phonetics and phonology (e. g., Erker and Guy 2012; Raymond and Brown 2012), and in areas as diverse as sociolinguistics (e. g., Grondelaers and Speelman 2007), historical linguistics (e. g., Gries and Hilpert 2010; Szmrecsanyi 2013; Wolk et al. 2013), and language acquisition (e. g., Ambridge et al. 2012). The majority of the above studies fall into one of two categories – those that use classification models to analyze corpus data and those that use classification models to analyze experimental data. However, our aim is to focus specifically on studies that combine textual and behavioral evidence. We will return to this issue in Section 2 where we define the scope of our survey.

While there are a number of diagnostics available to evaluate the goodness of fit and other properties of a model, the question of cognitive plausibility remains, i. e., how well do corpus-based models perform compared to native speakers, and how well do they capture what actually goes on in language. If we have a model with a classification accuracy significantly better than chance, are we to conclude that we have a good model? How should we define "good"? Occam's Razor encourages parsimony,[1] but what if the price we pay for parsimony is cognitive or linguistic plausibility? Is a corpus- or behavior-based model with high predictive and explanatory power satisfactory or should it be tested against native speakers' performance, viz. data on language in use? How can we determine whether the model says something about the cognitive processes behind the language use it aims to capture? Comparable performance can in theory be obtained by several entirely different processes.

In the present paper, we take a look at these and other pertinent questions, specifically as they regard the cognitive plausibility of statistical classification models. We present an overview of a number of behavioral studies that have been conducted to compare the results of corpus-based models for a range of lexical and syntactic phenomena in different languages. Our aims are to summarize previous research, present examples of good practice, draw attention to what works and what does not work, propose a methodological blueprint for future research, and open up the discussion of whether "validation" is the right way of looking at the phenomenon.

1 In statistical modeling, the principle of parsimony refers to the concept that a model should be as simple as possible. According to William of Occam, a fourteenth-century English philosopher, "the correct explanation is the simplest explanation" (Crawley 2007: 325).

## 2 Scope of the survey

There are by now many published corpus-based analyses of linguistic phenomena. In this paper, we focus on (1) multivariate corpus-based analyses using (2) data from a range of languages, including less widely studied languages, which have been (3) modeled statistically and compared or contrasted with (4) behavioral data obtained in experiments. The rationale behind these four criteria appeals to the fact that human language is inherently multivariate – although looking at a limited number of parameters may be desirable from a methodological point of view, it does not allow us to capture and study the phenomenon in all its natural complexity. For now, and for us, the gold standard is therefore a multivariate statistical analysis of an extensively annotated dataset: it captures more of the richness of a linguistic phenomenon and makes it possible for linguists to develop a multidimensional understanding. Looking at results from languages other than English instills us with confidence that the findings are not restricted to one language but will apply cross-linguistically. In this section, we define our selection requirements in more detail.

### 2.1 What do we include?

Complex, multivariate corpus models rely on a multitude of parameters to capture the essence of a linguistic phenomenon. One drawback of such datasets is that they are too large and complex for a human analyst to detect patterns without the aid of statistical methods. There are now plenty of published multivariate models that use data, extracted from corpora and annotated for a multitude of morphological, syntactic, semantic, and pragmatic parameters, to predict the choice for one morpheme, word, or construction over another. Prototypical examples are Gries (2003), Bresnan et al. (2007), De Sutter et al. (2008), and Klavan (2012) for a binary choice, Arppe (2008) for a four-way choice, and Divjak (2010) for a six-way choice.

Since, as cognitive linguists, we should ultimately be interested in the cognitive plausibility of corpus-based models, we insist on behavioral data from native speakers. However, only a small number of multivariate corpus studies have compared their findings with behavioral data (Wasow and Arnold 2003; Roland et al. 2006). Few have used authentic corpus data for this purpose (Arppe and Järvikivi 2007; Divjak and Gries 2008) and even fewer have directly evaluated the prediction accuracy of a complex, multivariate corpus-based model on humans using authentic corpus sentences.

The studies that meet the four requirements described above are Bresnan (2007; also Bresnan et al. 2007, Bresnan and Ford 2010; for a survey paper of

their work, see Ford et al. 2013a), Divjak (2010, Divjak et al. 2016b; also Divjak and Gries 2006, Divjak and Gries 2008, Divjak and Arppe 2013), Arppe and Abdulrahim (2013; also Arppe 2008), and Klavan (2014). While the studies by Bresnan and collaborators and Klavan are concerned with syntactic alternations, the studies by Divjak and collaborators and Arppe and Abdulrahim analyze lexical variation. The studies represent a range of European (English, Estonian, and Russian) and non-European languages (Arabic). Crosslinguistic evidence always increases the confidence we can have in our findings, and this is no different when it comes to methodology. After all, tools in the general linguistic toolbox should not be so specific as to be applicable to one language or language family only.

## 2.2 What falls outside the scope of this survey article?

Given our selection criteria, there are a number of studies that fall outside the scope of this survey article. Roughly, these fall into three categories: (1) studies that use statistical (especially logistic regression and related) modeling but do not combine textual and behavioral data; (2) studies that combine textual and behavioral data but limit the information that can be extracted from a corpus to token frequencies; and (3) studies that evaluate models arrived at using different modeling techniques by comparing their performance on one and the same dataset against each other.

When we look at studies that use statistical (logistic regression) modeling (and do it well), but only use one set of data, two distinct subgroups appear: studies that analyze experimental data statistically and studies that model corpus data statistically. In the first group, prime examples of work in linguistics that has been instrumental in developing Cognitive Linguistics were carried out in the MPI laboratories in Leipzig and Manchester, led by Michael Tomasello and Elena Lieven, respectively. Unlike in linguistics, statistical data analysis is a *sine qua non* in the field of psychology, which abounds with studies that employ statistical analysis. Yet, due to the nature of behavioral data (typically obtained as the result of a balanced experimental design), ANOVA long remained the most popular statistical technique. More recently, compelling arguments have been made to move toward (multilevel) mixed-effects models (Baayen et al. 2008; Jaeger 2008). As to the second group,[2] over the past fifteen years,

**2** This second group also comprises computational experiments or simulations that tend to focus on evaluating a trained model on unseen test data and do not typically involve human participants. Much of this type of research is carried out in the field of computational linguistics and natural language processing (see, e. g., Resnik and Lin 2010).

quantitative techniques have gained hugely in popularity within linguistics in general and within usage-based corpus-linguistic approaches in particular. Much work in the area has been done by Dirk Geeraerts' research unit "Quantitative Lexicology and Variational Linguistics" (QLVL) at KU Leuven, Stefan Th. Gries (University of California at Santa Barbara), and more recently by the "Cognitive Linguistics: Empirical Approaches to Russian" (CLEAR) group led by Laura Janda and Tore Nesset at Tromsø University. A range of statistical techniques have been introduced to the discipline; for an overview of some of the developments and main players in the field, see Glynn and Fischer (2010) and Glynn and Robinson (2014).

There are, of course, other ways in which corpus data and experimental data can successfully be combined. Gilquin and Gries (2009) give a – at the time – comprehensive overview of such studies, both from the perspective of psycholinguistics and of corpus linguistics. More examples are collected in Gries and Divjak (2012) and Divjak and Gries (2012). The most important variations on this theme – which are doubtlessly more frequent than the combination we advocate in this paper – are the following: relying on data extracted from corpora to design experimental stimuli; relying on frequency information calculated on the basis of data from a large and representative corpus to create frequency lists that are used to match items used in an experiment (see, e. g., Bradshaw 1984 for a catalog of studies published after 1960 that provide norms of material for use in experiments); including information extracted from corpora that is based on counting tokens (e. g., the frequency of occurrence of a form such as an ending/word/construction, n-gram specifications, morphological family size, etc.) when analyzing or modeling results of experiments (see Jurafsky 2003: 41–63 for an overview of earlier works in psycholinguistics). Some papers combine two out of the three options listed above. For example, Gries et al. (2010), van de Weijer et al. (2012), and Bermel and Knittl (2012a, 2012b) include information extracted from corpora (frequency counts, especially co-occurrence information) and use (fragments of) authentic sentences attested in the corpus as stimuli in their experiments, but they do not annotate the corpus data linguistically. Others, such as Bybee and Eddington (2006), Arppe and Järvikivi (2007), or Caines (2012), contrast textual and behavioral data but consider only one variable, i. e., a semantic similarity classification of the adjective that co-occurrs with one of four Spanish verbs of "becoming" in Bybee and Eddington (2006) or a classification of the subject in Arppe and Järvikivi (2007) and Caines (2012).

The third group consists of papers that compare the performance of different types of (statistical) modeling techniques on one and the same dataset; this is done in order to evaluate the performance of a statistical

model. Baayen et al. (2013), for example, compare the performance of logistic regression against the models of "tree and forest" and "naive discriminative learning" (NDL) on the basis of four datasets concerning rival forms in Russian. Their basic finding is that "the three models generally provide converging analyses, with complementary advantages" (Baayen et al. 2013: 253). Theijssen et al. (2013) evaluate their logistic regression model by comparing it to models arrived at using Bayesian Networks and Memory-based learning. Similarly to Baayen et al. (2013), the main finding of Theijssen et al. (2013) is that the performance of the three approaches is very similar. Baayen (2011) compares the performance of NDL with memory-based learning, logistic mixed-effects regression, and a support vector machine with a linear kernel to model the dative alternation in English. The conclusion Baayen (2011) draws is that the classification accuracy of NDL is on a par with the other classifiers; it is outperformed only by the support vector machine.

At the same time, Baayen's (2011: 311) comparison shows that the importance of individual predictors is evaluated differently by the different models. This means that although the performance of the different statistical models is comparable overall, they assign predictors different explanatory relevance. This is an oft-ignored yet crucial methodological point that brings us back to the thorny issue of how to decide which of the possible and probable models is the cognitively most plausible one. While the strategy of comparing the performance of different statistical models against each other certainly yields interesting insights, computational modeling cannot (yet) replace experiments involving native speakers. If the ultimate question we are trying to answer is "how to understand the statistical results from a cognitive perspective" (Baayen and Arppe 2011: 8), it is crucial, in our opinion, to compare the performance of any statistical analysis method used to model linguistic phenomena to the performance of native speakers in a controlled setting. This strategy is also advocated by Baayen, who admits that whether NDL or other models considered in the literature as cognitively more realistic are in fact closer to the truth "awaits further validation, perhaps through psycholinguistic experimentation" (2011: 311).

## 2.3 Why should you bother?

Now, why should (corpus) linguists run psycholinguistic experiments and why should cognitive scientists or psychologists interested in language conduct corpus-based analyses? Different from the contributors to the 2007 special issue of *Corpus Linguistics and Linguistic Theory* (volume 3, issue 1), we will

argue that a multivariate analysis of corpus data should be the gold standard in the discipline. Although we accept that there may be valid and pressing reasons to disregard the advice we present in this section, we argue that this disregard can only apply in very specific circumstances, and it should not be the default approach to an empirical study.

So far, we have avoided naming this procedure. Linguists who run an experimental study after a corpus-based study often refer to this process as "validation". This, unfortunately, creates the impression that behavioral data are inherently more valuable than textual data, be it transcribed spoken language or originally written language. But for language, textual data are the result of one of the most natural types of linguistic behavior, and "observing" the output qualifies as an "observational study" or a "natural experiment", which are quite popular in disciplines where experimental manipulation of groups and treatments would be unethical. Through observation, we get a real picture of the phenomenon as it manifests itself in natural settings.

In a natural setting, so many factors may influence a phenomenon that it becomes difficult if not impossible to establish cause-and-effect; although this can be (partially) countered by taking a multivariate approach, at present, we lack an exhaustive list of potentially influential factors, and we need experiments to check whether the factors we have identified by modeling observational data do indeed cause a particular behavior. It is also well known that there is a greater risk of selection bias in observational studies than in experimental studies, and this is certainly true for most studies based on small and medium-sized corpora; billion-word corpora may overcome this problem simply by being very large. A third set of issues relates to the observer him/herself: although observer bias, where the observer's interests color his/her observations, may well affect the annotations that are added to the corpus data, observer effect tends to manifest itself mainly when the data consist of transcribed recordings of conversations, as in acquisition corpora where a researcher was present during the recorded interactions. And finally, not every pattern that we can detect in a large dataset will have been picked up by every speaker, and this problem is only becoming more acute as the size of datasets increases and statistical modeling techniques become more sensitive. We need experiments – at least for now – to set upper and lower boundaries to what could be psychologically relevant and to calibrate our models.

Some of the reasons listed above also explain why psycholinguists should consult a corpus and not routinely limit their interest to a unit's frequency of occurrence. For one, in a natural setting, so many factors influence the phenomenon that selecting factors without an exhaustive, i. e., multivariate, study of the phenomenon equals trying one's luck. Considering frequency counts alone

severely impoverishes the richness of the linguistic experience from which learners extract distributional patterns. Second, experimental studies often use artificially constructed stimuli that bear little resemblance to naturally produced data. They thus "force participants to tackle problems that are not faced in normal discourse" (Deignan 2005: 117) and are therefore of limited use for validating predictions of linguistic theories. In fact, the external validity of many laboratory results has been questioned. Mitchell (2012) aggregated results of several meta-analyses and concluded that, although many psychological results found in the laboratory can be replicated in the field, their effects often differ greatly in size and sometimes even in direction (Mitchell 2012: 114); developmental work fared exceptionally poorly in this respect, showing a large negative correlation between lab and field results. Mitchell (2012: 115) also stresses that alternative divisions of the data would yield different patterns. These shortcomings of lab-based research can be addressed by statistically modeling multivariate corpus data as this allows us to study and assess the effect of a variable in competition with a multitude of others.

Dealing with the effects of frequencies in language use on cognition is a discipline at the intersection of cognitive corpus linguistics and psycholinguistics (see Divjak 2012: 3). As Chafe (1992: 96) put it:

> But I continue to believe that one should not characterize linguists, or researchers of any kind, in terms of single favorite tie to reality. The term "corpus linguist" puts the emphasis on one tie to reality that has been neglected by many contemporary linguists, I believe to the great detriment of the field: a tie that must be vigorously pursued if our understanding of language and the mind is to enjoy significant progress. But there is a complementary danger in implying that that is all a linguist should do, of pitting corpus linguists against introspective linguists or experimental linguists or computational linguists. I would like to see the day when we all be more versatile in our methodologies, skilled at integrating all the techniques we will be able to discover for understanding this most basic, most fascinating, but also most elusive manifestation of the human mind.

Being "skilled at integrating all the techniques" (McEnery and Hardie 2012: 226) may be too much to ask – is it really feasible for a linguist to be an expert in multivariate corpus analysis, statistical modeling, and experimental design, including different off-line and online paradigms? One thing that the studies surveyed in this article have in common is their "longevity"; the topics have been looked at from various angles over a period that can easily span ten years. Change takes time: methods need to be identified, skills need to be acquired and honed, often through trial-and-error. Although subsequent studies are more straightforward and less time consuming to run, some techniques such as fMRI are (and will be for the foreseeable future) too expensive to be widely

used. Linguistics as a field may need to embrace collaboration across disciplines, which presupposes a basic knowledge and understanding of cognate disciplines, a convergence in research methodology, and specific, testable hypotheses.

## 3 Probabilistic statistical classification models

Classification is a vital process, both from a cognitive and mathematical perspective. In cognitive and linguistic terms, we often refer to classification as categorization.[3] Mathematical, or more precisely, statistical classification involves identifying to which category a new observation belongs on the basis of observations whose category has already been established, i. e., on the basis of a training set. Classification is a common "problem" in many areas of scientific research, and there is a vast array of statistical methods available for solving such "problems", e. g., cluster analysis, discriminant analysis, (logistic) regression, tree-structured methods such as CART, "tree and forest" methods, memory-based learning, etc.

Our focus is on a subclass of classification, namely probabilistic classification, where statistical inference is used to find or predict the best class for as yet unobserved "observations". This is done by calculating the probability of an observation belonging to a set of possible classes. Statistical models can be deterministic (e. g., physical laws) or they can be highly indeterminate predictive models with large predictive errors (e. g., a model representing attitudes as a function of socioeconomic factors) (Kotz 2006: 7080). It would be fairly safe to assume that models representing the complex nature of natural language data tend to fall into the latter category. Linguistic data can be described as "messy data" since there is a lot of intercorrelation among the explanatory variables; this is also referred to as "rampant collinearity" – something that can pose serious problems for statistical analysis, certainly if it is not also present in the larger population from which the data are drawn (see Harrell 2001: 65). Moreover, in linguistic analysis, the number and levels

**3** The mathematical details of logistic (regression) modeling and other modeling techniques fall out of the scope of this paper; the interested reader is referred to the accounts given, for example, in Crawley (2007: Chapters 9, 10, 13, 18, 19, 20, and 21), Baayen (2008: Chapters 6 and 7), and Hosmer et al. (2013).

of explanatory variables are large. All of this entails that much thought must go into choosing the relevant explanatory variables in order to determine the "best" predictive model.

Model selection – deciding which variables to include – is a crucial step in statistical modeling and, unfortunately, not a straightforward one (see, e. g., Harrell 2001; Burnham and Anderson 2002; Hosmer et al. 2013). There are a number of strategies for arriving at the best model, and opinions are divided as to which is best. Common sense dictates that it is our theory that should guide us as to which variables to include, and not the *p* values that accompany the variables in a full model. In any case, it should be borne in mind that there is not one sacred model that can be fitted to a particular dataset. It is far more likely that there will be a number of different plausible models that fit the data. In fact, multimodel inference techniques (e. g., Burnham and Anderson 2002) are emerging within linguistics (e. g., Barth and Kapatsinski in press). Faced with a situation where a number of models fit the data equally well, we once again return to the question of which of the alternative models is cognitively most plausible. Behavioral data may offer a helping hand here – comparing the performance of alternative models to the performance of native speakers may guide us when determining which of the possible models is closest to the elusive "truth". Experiments also prove useful when we need to verify the importance of individual factors and tease apart the complex structure in a controlled "laboratory" setting (but see also the discussion above in Section 2.3).

Following model building, the goodness of fit of the model needs to be assessed. A model's goodness of fit refers to the knowledge of "whether the probabilities produced by the model accurately reflect the true outcome experience in the data" (Hosmer et al. 2013: 153). In a very basic sense, measures of goodness of fit summarize the agreement between observed and fitted values. Or to put it another way, they measure the difference between the observed and fitted values. A question that could be expected to arise naturally in the context of a discussion about the classification accuracy of a statistical model, but that has been often overlooked, is the question of how good human classification is? For any model, we fit to corpus data, cognitive linguists should be interested in finding out whether human classification is better, worse, or on a par with that of a statistical model. The less cognitively inclined can think of native speaker performance as an additional model validation technique. Several suggestions have been made in the literature concerning the comparison between "men and machines": (1) human classification seems to be outperformed by some machine classifiers when generalizing to unseen data (Baayen 2011: 306); (2) different from machine learning, human learning might be (more) susceptible to variation (e. g., individual speaker variation) (Baayen 2011: 313; Divjak et al. 2016b);

(3) there are inherent differences in how a statistical model derives a quantitative structure of distributional patterns and how a human being acquires this knowledge (Baayen 2011: 317; Milin et al. 2016).

In order to determine whether statistical classification is comparable to human classification, we need behavioral data. Three logical conclusions follow from comparing our model fitted to the corpus data to native speaker performance in a linguistic experiment (see Divjak et al. 2016b): (1) if human performance is on a par with that of our fitted model, we can add certainty to our conclusion that the model we have selected "has a good fit" and is "cognitively plausible"; (2) if human performance is inferior to the model, we may suspect that the model is more complex than the actual reality; (3) if human performance is superior to the model, we are most likely missing some important predictors from our model formula. We now turn to this discussion by presenting an overview of the four corpus and experimental studies we consider as test cases in this paper.

# 4 Corpus models included in our overview

The four case studies we concentrate on are the English dative alternation (Bresnan 2007; Bresnan et al. 2007; Bresnan and Ford 2010; Ford et al. 2013b), the alternation between the adessive case suffix and the postposition *peal* in Estonian (Klavan 2012, 2014), six Russian verbs denoting the concept of "try" (Divjak 2003, 2004, 2010; Divjak and Arppe 2013; Divjak et al. 2016b), and four near-synonymous verbs meaning "come" in Modern Standard Arabic (Abdulrahim 2013; Arppe and Abdulrahim 2013).

## 4.1 Corpus data

Table 1 presents an overview of the corpus data used in the four studies; it specifies which corpora were used and whether the data come from written texts (the Estonian and Russian studies), spoken language (the English study), or both (the Arabic study). In the third column, the size of the database is given, with the number of tokens for a particular construction or lexeme in parentheses. The fourth column shows how many different properties were annotated. We will not report the complete annotation schemes used in these studies but refer instead to the literature for details. Suffice it to say that, in general, the majority of the properties pertain to discourse semantics, syntax, and morphology.

**Table 1:** Overview of the corpus data used in the four studies.

| Case study | Corpus | Number of observations | Properties |
|---|---|---|---|
| English: NP NP vs. PP[a] | *Switchboard Corpus* (telephone conversations) | 2,360 (1,859 NP NP, 501 PP) | 14 (29 levels + 1 interval variable) |
| Estonian: adessive vs. *peal* | *Corpus of Written Estonian* (fiction and newspaper texts 1980–2000) | 900 (450 per construction) | 20 (46 levels + 1 interval variable) |
| Russian: 6 "try" verbs | *Amsterdam Corpus* and *Russian National Corpus* (written literary texts 1950–2000) | 1,351 (~250 per verb) | 14 (87 levels) |
| Arabic: 4 "come" verbs | Modern Standard Arabic component of *arabiCorpus* (written and spoken language) | 2,000 (500 per verb) | 22 (72 levels) |

Note: [a]NP NP refers to the double object construction and PP to the prepositional dative.

## 4.2 Model building and evaluation of fit

### 4.2.1 The English dative alternation

Bresnan et al. (2007) employ both a simple binary logistic regression and a more complex regression technique – often called "mixed-effects model" (Pinheiro and Bates 2000) – to model the response, the choice between the dative NP or the dative PP, as depending on a range of predictor variables. They present three models fitted to the data. We focus on the model that is used in the experimental cross-validation study (Bresnan 2007), referred to as Model B in the original paper. This model uses "verb senses" as the random effect in addition to the fixed effects. This means that the binary response is conditioned by the 55 verb senses present in the dataset. Their model formula has 14 predictors (given in the order they appear in the original formula): semantic class of the verb used in the construction (5 classes in total), accessibility of recipient and theme (given vs. not given), pronominality of recipient and theme (pronoun vs. non-pronoun), definiteness of recipient and theme (definite vs. indefinite), animacy of recipient (animate vs. inanimate), person of recipient, number of recipient and theme (singular vs. plural), concreteness of theme, structural parallelism in dialogue (existence of the same kind of structure in the same dialogue), and length difference (the difference in

number of graphemic words between the theme and recipient, taking a sign-preserving log transform of the absolute value of the difference to reduce the effect of outliers).

When building their model, Bresnan et al. (2007) chose to include all the variables that were considered at the annotation stage. They included factors that did not have a statistically significant effect on the outcome, the reason being that "[e]liminating such variables often biases the results by inflating the apparent magnitudes of the effects of other variables (Harrell 2001)" (Bresnan et al. 2007: 83, footnote 3). Their model correctly classifies 95 % of the data overall. The average percentage of correct predictions when testing the model on data from the same corpus is 94 %. This measure was obtained by randomly dividing the data 100 times into a training set, fitting the model parameters on each training set ($n = 2{,}000$), and scoring its predictions on the unseen testing set ($n = 360$).

#### 4.2.2 The Estonian adessive case vs. postposition *peal*

Klavan (2012) used a dataset of 450 tokens of the adessive case and 450 tokens of postposition *peal* to train a binary logistic regression to predict the choice between the two constructions in present-day written Estonian. The original dataset contained one interval variable and 19 categorical variables with 46 distinct variable categories or contextual properties. A stepwise model simplification strategy was adopted (Crawley 2007: 323–386), i. e., a minimally adequate model was selected from a large set of more complex models on the basis of deletion tests (*F*-tests or Chi-squared tests) which assess the significance of the increase in deviance that results when a given term is removed from the model. An explanatory variable was only retained in a model if it significantly improved the fit of the model, i. e., if it caused a significant increase in deviance when removed from the model. Altogether six predictors were retained in the final model: mobility of the landmark (LM) phrase (mobile vs. static), verb group (action, existence, motion, posture, and no verb), length of LM phrase in syllables (log. transformed), morphological complexity of LM phrase (compound vs. simple), word class of trajector (TR) phrase (noun, pronoun, and verb phrase), and relative position between TR and LM (LM before TR, TR before LM).

The final model surpasses competing models in its goodness-of-fit statistics, predictive power, accuracy of prediction, and diagnostic results. The model correctly classifies 70 % of the data overall. Although for a binary choice, 70 % as an overall accuracy of the model may seem low (especially when we compare

this model to the models fitted to the English dative alternation), the question arises whether this reflects the "reality" (i. e., it is in fact difficult to tease these two constructions apart, no matter which variables are modeled in) or whether the model can be substantially improved by adding crucial predictors. Comparing the performance of the corpus model to behavioral data may help in finding an answer to this question.

### 4.2.3 Russian verbs expressing "try"

Building on earlier work by Divjak (2003, 2004, 2010), Divjak and Arppe (2013) used a dataset of 1,351 tokens to train a polytomous logistic regression model, applying the one-vs.-all heuristic (Arppe 2013a, 2013b) to predict the choice between 6 Russian verbs that, when combined with an infinitive, can all be translated with the English verb *try*. At the model building stage, a number of criteria were followed: the number of variable categories in the model was not to exceed 1/10 of the least frequent outcome,[4] only variables with a broad dispersion among the six verbs were to be retained (the overall frequency of the variable in the data was to be at least 45 and to occur at least twice with each verb), one variable was to be excluded for each fully mutually complementary case, as were variables with a mutual pair-wise uncertainty coefficient value larger than 0.5 (Divjak and Arppe 2013: 238). Eighteen predictors (11 semantic and 7 structural) were retained in the model, belonging to seven different variables encoding the TAM (tense–aspect–mood) marking on the verb, the semantics of the subject and infinitive, and properties of clause and sentence. The model predicts the probability for each verb in each sentence and reveals how strongly each feature individually is associated with each verb (Divjak and Arppe 2013: 236–237). The model's overall accuracy is reported as 51.7 % (50.3 % when tested on unseen data). Divjak and Arppe (2013: 242) point out that although this may seem low, it is well-above chance for a six-way choice (chance performance would be 16.7 %) between options that display the exact same constructional possibilities and limitations. Furthermore, the more interesting question is how the model's performance compares with humans – something we will discuss in Section 5.

4 Arppe (2008: 116) points out that the number of distinct variable combinations that allow for a reliable fitting of a (polytomous) logistic regression model should not exceed 1/10 of the least frequent outcome. In the Russian dataset, the least frequent verb occurs about 150 times – hence the selection criteria of 15 variable categories.

### 4.2.4 Arabic verbs meaning "come"

Arppe and Abdulrahim (2013; see also Abdulrahim 2013) fitted a polytomous logistic regression model based on the one-vs.-all heuristic (Arppe 2008, 2013a) to 2,000 sentences retrieved from *arabiCorpus*. Its aim was to determine the relative effects of multiple predictor variables on the choice between four "come" verbs in Modern Standard Arabic that are considered near-synonyms in some dictionaries (see Arppe 2008; Divjak 2010). The annotation schema included 22 morpho-syntactic and semantic properties or contextual features (Abdulrahim 2013: 46; Arppe and Abdulrahim 2013). The following steps were taken when building the model and selecting the set of predictor variables:[5] (1) inspection of the distribution of variables across all "come" verbs using standardized Pearson residuals (variables with a value approaching 0 would not be included in the polytomous logistic regression model); (2) inspection of pair-wise association patterns between variables (only one of two highly associated variables was selected for the model; for variables that are symmetrically complementary, only one was included in the model); (3) selecting only predictor variables with an overall frequency of 20 and with at least 10 occurrences for two verbs (Abdulrahim 2013: 63). The final polytomous logistic regression model fitted to the data includes 31 predictor variables (Abdulrahim 2013: 162–163), which pertain to (1) the morphological properties of aspect, mood, tense, and the gender, number, and person of the subject; (2) syntactic properties of adverbial phrase, locative adverb phrase, prepositional phrase, transitivity, and negation; and (3) semantic properties of the semantic category of subject and different types of phrase (comitative, goal, manner, purposive, setting, source, and temporal phrase). The model accuracy is reported as 0.845 (Abdulrahim 2013: 164; Arppe and Abdulrahim 2013). A psycholinguistic experiment was run to compare the probability estimates calculated by the model with lexical choices made by speakers of Arabic (Arppe and Abdulrahim 2013, see Section 5).

### 4.2.5 Classification accuracy of the corpus models

Table 2 summarizes the classification accuracy measures for all four datasets. The second column in this table indicates how many values the dependent or

5 Prior to polytomous logistic regression analysis, the nominal form of the Arabic data frame was converted to a logical one, whereby every level of each of the original 22 variables was turned into a variable in its own right, with the binary values TRUE/FALSE; for example, the binary levels YES/NO for the variable GOAL were converted into two variables: GOAL.YES and GOAL.NO (Abdulrahim 2013: 161).

**Table 2:** Classification accuracy of the four corpus datasets, sorted by rate of improvement.

| Language | Outcome levels | Predictors | Chance correct (%) | Baseline correct (%) | Model accuracy (%) | Improvement model over baseline |
|---|---|---|---|---|---|---|
| English | 2 | 14 | 50 | 78.7 | 95 | 1.2 |
| Estonian | 2 | 6 | 50 | 50 | 70 | 1.4 |
| Russian | 6 | 18 | 16.7 | 18.5 | 52 | 2.8 |
| Arabic | 4 | 31 | 25 | 25 | 85 | 3.4 |

outcome variable has. The third column specifies how many predictors were included in the model. The column "model accuracy" gives the proportion of correctly classified data – we can see that for all four models, the classification accuracy is significantly better than chance (fourth column) and better than baseline (fifth column). Prediction accuracy should not be taken at face value, however. For instance, in the English dataset, there were a significantly higher number of double object constructions in the dataset than prepositional datives (1,859 instances of NP NP construction and 501 instances of PP construction). So, the actual chance of being right is higher than 50 % (about 79 %) if the most frequent option is always chosen. This "improvement" rate is given in the seventh column and was calculated by dividing the accuracy by baseline. That being said, always selecting the most frequent outcome is intelligent behavior, and in line with the foundational assumptions of usage-based linguistics.

But is "significantly better than chance" good enough? Assuming that other model-fitting diagnostics are also "in good order", are we then to conclude that we have, indeed, a good model? In the next section, we will argue that model performance needs to be measured against human performance and presents a number of ways in which this has been done.

# 5 Experimental validation studies of corpus models

In this section, we report on the findings of experimental studies which have been run to compare and validate a statistical model fitted to corpus data (Bresnan 2007; Bresnan and Ford 2010; Ford et al. 2013b; Arppe and Abdulrahim 2013; Klavan 2014; Divjak et al. 2016b). We focus on studies that have validated the corpus-based

model in its entirety, rather than focusing on the cognitive reality of one particular (set of) predictor(s).[6]

## 5.1 Experimental designs, materials, and participants

For the Russian (Divjak et al. 2016b), Estonian (Klavan 2014), and Arabic (Arppe and Abdulrahim 2013) studies, a forced choice task with a similar experimental design was carried out. The stimuli came from the original corpus studies: 60 sentences from Divjak's (2010) dataset, 30 sentences from Klavan's (2012) dataset, and 50 sentences from Abdulrahim's (2013) dataset. In all three studies, the selection of the stimuli was driven by the rationale that the sentences represent the diversity of the probability distributions. This means that the chosen stimuli ranged from sentences where one construction or verb was very probable (near-categorical preferences) to sentences where both constructions or all verbs were virtually equally probable (approximately equal probability estimates for both/all choices). Respecting the textual distribution also means that the verbs are not necessarily represented by the same number of stimuli in the experiment; for instance, the list of Russian stimuli contains 4, 6, 8, 10, 12, and 20 examples for each of the verbs. The general prediction of these studies is that the proportion of choices made by the native speakers mirrors the probabilities estimated by the statistical model. The corpus sentence was presented to the respondents with a blank for the original construction or verb; all (2, 4, or 6) options were presented. The respondents were asked to choose which of the alternative constructions or verbs fits the context best. For the Estonian and Russian study, four experimental lists were created, each with a different random order. Ninety-six native speakers of Estonian, 134 native speakers of Russian, and 30 speakers of Bahraini Arabic participated in the studies.

**6** Assuming that one possible scenario emerging from comparing corpus-based models to behavioral data is that the corpus model outperforms humans and that not all of the linguistic predictors included in the corpus model are picked up by the native speakers, we may want to run additional experimental studies in order to take a closer look at individual predictors. For example, Divjak et al. (2016a) focus on TAM-marking using a self-paced reading paradigm. For Estonian, Klavan (2012) found that the length of the phrase, which was the strongest predictor in the corpus-based model, has only a small significant effect in the opposite direction in an acceptability judgment study. The stimuli were not authentic corpus sentences, however, and hence, a direct comparison between the corpus and behavioral data should be approached with caution.

For the English dative alternation (Bresnan 2007, see also Bresnan and Ford 2010; Ford et al. 2013a, 2013b), a slightly modified version of the forced-choice task was carried out, referred to as a "forced choice scalar rating task" or the "100-split task" (Ford et al. 2013a, 2013b). In this task, the respondents had to rate the naturalness of the two constructional alternatives by distributing 100 rating points over them. The aim was to evaluate how well the naturalness ratings of these two alternative syntactic paraphrases correlated with the corpus probabilities estimated by the logistic regression models. Bresnan (2007) used 30 experimental items (authentic passages attested in a corpus of transcriptions of spoken dialogue) taken from the original corpus study by Bresnan et al. (2007). The passages were randomly sampled using the corpus model probabilities, ranging from a very low to a very high probability of having a prepositional dative construction. For each sampled observation, an alternative paraphrase was constructed, and both options were presented as choices in the original dialogue context. Nineteen native speakers of English participated in the Bresnan (2007) study with all participants receiving the same questionnaire, with the same order of items and construction choices.

## 5.2 Results

When comparing the performance of the corpus-based model with that of the native speakers, two different approaches were taken. Bresnan (2007) and Arppe and Abdulrahim (2013) analyzed the responses given by the participants as a function of the original corpus model predictor variables using mixed-effects logistic regression. The forced-choice selections or the scores in the scalar rating task were modeled as dependent variables with the original corpus model predictor variables as independent variables (fixed effects) and participants as random effect. Divjak et al. (2016b) propose a different approach, also adopted by Klavan (2014). For the Estonian and Russian data, the 30 and 60 sentences that were used in the forced-choice task were excluded from the dataset. The corpus-based model was then trained on the remaining sentences and used to compute the probability of the two constructions or six verbs in the sentences used for the experiment. This allows a direct comparison of how the corpus model performs compared to native speakers in predicting the choice in the specific 30 or 60 sentences.

### 5.2.1 Results of the Arabic study

Arppe and Abdulrahim (2013) compared the proportions of selected verbs for each of the 50 experimental items with the matching corpus-based

probability estimates and show that there is a high and significant correlation ($r_{Pearson} = 0.747$; $p < 2.2^{e-16}$) – as the probability of a verb rises, so does the proportion of selections of that verb. They also fit a mixed-effects logistic regression model with the individual forced-choice verb selections as the dependent variable, the context incorporated in the stimuli as independent fixed effects variables, and participant as random effect. As to the goodness of fit of the model ($R_L^2 = 0.312$ and Accuracy = 0.636), Arppe and Abdulrahim (2013) conclude that its performance can be considered very good: the estimated verb-specific odds in the mixed-effects model of behavioral data were found to largely agree in both direction and strength with those of the original corpus-based model. As the authors point out, there is validation for the selection of the explanatory variables used in the corpus-based model and further cross-linguistic corroboration of the hypotheses on the (positive and linear) relationship between corpus-based relative frequencies and proportions in forced choice, as also presented in Arppe and Järvikivi (2007).

### 5.2.2 Results of the English study

For the original scalar rating task, Bresnan (2007: 78) hypothesized that "given the same multivariable information as the corpus model, including contextual information from the original dialogues, subjects will make ratings of alternative dative constructions [...] that correspond to the corpus model probabilities." The results were analyzed using a linear mixed-effects regression model with subject and verb sense as random effects. An initial model was run with the fixed effects from the original corpus model of Bresnan et al. (2007) together with the order of items, the order of construction choice, and the lemma frequency of the verbs. The last three effects were eliminated from the final model because their coefficients were less than their standard errors. To assess the fit of the experimental model, Bresnan (2007) inspected the scatterplots where all 30 scores were plotted against the fitted model values of the data for each 19 subjects. She concluded that the model variables show a good fit to the behavioral data ($R^2 = 0.61$, Bresnan 2007: 84). Furthermore, the subjects' preferred choices reliably picked out the same choices made in the original corpus transcriptions. The mean proportion of subjects' ratings favoring actual corpus choices is 76 % (ranging from 63 % to 87 %) (Table 2 in Bresnan 2007: 84). The baseline was 57 %; this means that if the subjects had invariably preferred the double object construction in every experimental item, their responses would have matched 57 % of the original choices. The overall conclusion is that "subjects' scores of the naturalness of the alternative syntactic structures

correlate very well with the corpus model probabilities and can be substantially explained as a function of the same predictors as the original corpus model" (Bresnan 2007: 84).

### 5.2.3 Results of the Russian study

To obtain comparable data, Divjak et al. (2016b) first excluded the 60 test sentences from the original dataset (Divjak 2010) and trained the model on the remaining sentences. The probability for each of the six verbs in each of the test sentences was then computed using the new model. In order to compare the model probabilities and the choices made by participants in the forced choice task, it was assumed that the verb with the highest predicted probability for a given context would be the model's response on the same forced choice task. The "correct" response is taken to be the verb which actually occurred in the original corpus sentence. The model predicted the "correct" verb for 23 of the 60 test sentences (38 % of the time). Divjak et al. (2016b) point out that the testing set intentionally contained a larger proportion of verbs in highly variable contexts than would be the case in a random sample. The mean number of "correct" choices for the participants was 27.7 (46 % of the time, SD 4.7, median 28) and 21.2 when penalized for guessing. The scores ranged from 13 to 38, indicating considerable individual variation.

Overall, the authors conclude that although both model and speaker perform 2.5 to 3 times better than chance, they still make the "wrong" choice in more than half of all cases. In contrast to speakers, the model had no access to information about the token frequencies of individual verbs other than their frequencies in the corpus sample, which were roughly equal by design. It was therefore decided to accommodate frequency information into the model by multiplying the predictions of the original model by the square root of each verb's relative frequency. The frequency-adjusted model predicted the target verb correctly in 28 of 60 sentences, i. e., it performed exactly at the same level as the average human participant.

Divjak et al. (2016b) also conducted a second analysis to see how often the participants, the model, and the corpus "agreed" (i. e., both the model and the participants chose the verb that occurred in the corpus), and how often they "disagreed". The verb that was selected by the largest number of participants was deemed to be the preferred choice. There were 9 out of 60 (i. e., 15 % of all) experimental items where the model and the participants agreed and where the choice of the verb attested in the corpus was unusual or obsolete, and the verb preferred by the participants (and the model) should be regarded as "correct". Thus, the participants' (and the model's) true performance may be about 15 % better than reported above.

### 5.2.4 Results of the Estonian study

For the purpose of comparing the corpus model to native speakers, Klavan (2014) adopted the method of analysis proposed by Divjak et al. (2016b). First, the 30 test sentences from the original dataset were excluded, and the model was trained on the remaining sentences. It was assumed that the construction with the highest predicted probability would be the model's response on the forced choice task. The "correct" response was taken to be the construction actually used in the original corpus sentence. There were 30 sentences and 2 constructions in the task – chance performance would thus be 15/30. For the 30 sentences randomly sampled for the experiment, the model predicted as many as 27 sentences out of 30 correctly, yielding a prediction accuracy of 90 %. However, had a different set of 30 sentences been used, the prediction accuracy would have been lower. Averaged over 40 random subsets, the prediction accuracy was 71 % (comparable to the accuracy measures reported above for the full corpus model). The mean number of "correct" choices for the participants was 22.6 (accuracy 75 %, median 23, SD 2.5). Similarly to what Divjak et al. (2016b) observed in their behavioral data, there was also considerable individual variation among the Estonian speakers (with scores ranging from 14 to 28). If the results are corrected for guessing,[7] the average prediction accuracy is considerably lower (mean 15.2, SD 5.02, accuracy 50 %, median 16).

## 5.3 Summary and discussion of the case studies

Let us now look at the results of all four case studies in tandem and return to the question of cognitive plausibility. To recapitulate, there are three logical outcomes: human performance is on a par with, inferior, or superior to that of the fitted corpus model. Table 3 compares the accuracy measures of the corpus models to those of the native speakers in the experimental studies.

7 Within the fields of educational measurement and psychometrics, it is common to use a correction for guessing formula in order to account for random guessing during a multiple-choice task. Since we can never be sure what it is that speakers are actually doing during a linguistic task (i. e., are they making an intentional informed choice or just guessing), a similar approach can be taken in linguistics when analyzing behavioral data in order to increase the validity of the results. Klavan (2014) used a strategy referred to as formula scoring (Frary 1988) for correcting the results of her experimental study. The formula: $FS = R - W/(C - 1)$, where FS = "corrected" score, R = number of items answered right, W = number of items answered wrong, and C = number of choices per item.

**Table 3:** Agreement between the corpus models and native speakers in the four experimental case studies.

| Study | Predicting writers' choices | Predicting speakers' choices |
|---|---|---|
| Arabic | NA | Accuracy = 64 %<br>* Chance = 25 % |
| English | NA | Accuracy = 76 %<br>* Baseline = 57 % |
| Estonian | 27/30 = 90 % (21.3/30 = 71 %)[a]<br>* Chance = 15/30 (½) = 50 % | 22.6/30 = 75 %<br>* Baseline = 53 % |
| Russian | 23/60 = 38.3 %<br>* Weakest baseline = 4/60 = 0.6 %<br>* Best baseline = 20/60 (⅓) = 33.3 % | 27.7/60 = 46 %<br>* Weakest baseline chance = 10/60 (⅙) = 16.6 %<br>* Best baseline = 20/60 (⅓) = 33.3 % |

Note: [a]The measure in parentheses is the average prediction accuracy computed for 40 random subsets of 30 items.

It should be borne in mind that the four studies differ on a number of dimensions, of which we will recapitulate the most important ones here. First, there were two different approaches to analyzing the behavioral data when comparing the behavioral data to the corpus data: while the approach taken by Bresnan (2007) and Arppe and Abdulrahim (2013) allows to model the responses given by the participants as a function of the original corpus model predictor variables using mixed-effects logistic regression, Klavan (2014) and Divjak et al. (2016b) compare how the corpus model performs compared to native speakers in predicting the choice in a specific subset of sentences. Second, the number of participants in the experiments differed greatly, with 19 and 30 in the English and Arabic studies vs. 96 and 134 in the Estonian and Russian studies (see the discussion in Divjak et al. 2016b on the discrepancy between individual and group performance). Third, the degree of "synonymy" between the alternatives was different for each case study, as was the number of synonyms included (two vs. four or six). It is only to be expected that choosing between six alternatives as opposed to say four or two alternatives will be more demanding for both the statistical corpus model as well as native speakers. Moreover, not all choices are created equal: the more similar the items are, the harder it will be to distinguish between them. Given that linguists do not apply the same criteria when selecting near-synonyms (see Divjak 2010: 105 for a discussion), some choices may be trickier than others, and this will influence the prediction accuracy of both model and speakers.

If we were to consider the corpus model only, we would probably conclude that the Estonian and Russian models are not doing a particularly good job. For the full dataset, the binary Estonian model had an accuracy of 70 % and the Russian six-way model 52 %. However, when comparing how humans perform on the same set of data, it becomes clear that both the Estonian and Russian models perform at more or less the same level of accuracy as native speakers; this suggests that the corpus-based models are cognitively plausible. For the Arabic and English datasets, human performance seems somewhat inferior to that of the corpus-based model. Given the accuracy measures in Table 2 and Table 3, we can see that the Arabic and English corpus models outperform humans (85 % vs. 64 % and 95 % vs. 76 %), suggesting the truthfulness of the maxim "not everything that can be counted counts". It would appear, then, that there are variables in the corpus model that the native speakers are not picking up on. Both corpus models had, in fact, a large number of predictors (14 in the English study and 31 in the Arabic study). It may therefore be reasonable to assume that while all of these predictors combined yield statistical models with excellent fit, human classification is less precise, and we may need to take this into account when we calibrate our models. A good model, like a good theory, should be parsimonious, precise, and testable. It is our job as researchers to decide where to strike the balance and combining textual and behavioral data should help us do this. Divjak et al. (2016b) feed insights gained from the experiment back into the model: once the original model was adjusted for frequency, it performed exactly at the same level as the average human participant. The model started to behave more like speakers with overgeneralizations of the most frequent verb. This finding provides support for the assumption that speakers do use frequency information and that, if possible, it should be included in the corpus-based model.[8] As for the Estonian data, we can also say that for the 30 test items, the corpus-based model outperformed humans (90 % vs. 75 %). However, a more realistic estimate for the accuracy of the corpus model is probably around 70 %, which is the accuracy for the full model and for the 40 random subsets of 30 items.

Overall, the authors of all four studies conclude that the results of the corpus-based model and behavioral data, by and large, converge. In other

**8** In this particular case, the unigram frequency of the six verbs differs so much (ranging from <0.01 for the least frequent verb to 0.56 ipm for the most frequent verb) that it was not feasible to respect the natural proportions while including a sufficiently large sample of examples for the least frequent verb.

words, participant's responses correlate well with the corpus model probabilities, allowing us to draw the conclusion that there must be some cognitive reality to the corpus-based models. Of course, it is also possible that a different set of predictors will yield alternative corpus models with accuracy measures comparable to human beings, but both Bresnan (2007) and Arppe and Abdulrahim (2013) were able to show that subjects' responses can be explained as a function of the same predictors as the original corpus model. Yet, there is some evidence (Theijssen et al. 2013: 258) that higher level features, such as syntactic, semantic, and discourse-related features, do not dramatically improve model performance when compared to models with lexical features alone (i. e., the actual words used). This calls into question the assumption that humans make use of such abstract higher level features when choosing between alternatives and calls for further empirical research.

## 6 Where do we go from here?

We hope to have shown that addressing the topic of this special issue – how empirical results feed back into theory – sends us back to square one as it requires considering which kind of data we base our analyses on and how these data types relate and interact. As a blueprint for future research, we would like to emphasize methodological pluralism: what we need for the field to move forward is multivariate corpus research coupled with experimentation. Researchers often settle for a set of parameters mentioned in previous studies, but very few of these have been tested on actual speakers or actual collections of data in use. As a result, we (still) do not know with certainty what parameters are or could be important – we therefore need to cast the net wide and test complete corpus models as well as individual variables on speakers, bearing in mind that our behavioral findings may well necessitate changes to our corpus models.

In this contribution, we have shown that many linguistic phenomena are not fully predictable. Relying solely on corpus models may give a false impression of the corpus models themselves – sometimes models are very accurate, and sometimes they appear to be less accurate, but if interest is in modeling human behavior, then we need behavioral data to evaluate the model. Although the field is beginning to embrace more intricate modeling techniques (e. g., memory-based learning and NDL) which seem cognitively more plausible, considering behavioral evidence remains a *sine qua non*. It is hoped that if the field of linguistics embraces the approach advocated in the paper, testing corpus-based models against human performance will become one of the goodness-of-fit criteria for assessing model performance.

One of the objections frequently heard when corpus data are compared to behavioral data is that this approach compares apples to oranges. One set of data is said to reflect production and the other comprehension. The picture is far from clear-cut: written language is hardly pure production and no comprehension (we often reread and edit a text), just as choosing between alternatives is hardly all comprehension and no production (see Tooley and Bock 2014 and references therein for a recent overview on the relationship between language comprehension and language production). Arppe and Abdulrahim (2013) have even claimed that making a selection can be seen as a form of production that is comparable to the process underlying the generation of corpus data. Moreover, Bresnan and Ford (2010) show that their corpus-based models accurately predict behavior for rating, comprehension, and production.

Another objection concerns the type of experiments that would adequately complement a corpus-based model. In the four studies we discussed, off-line forced-choice and acceptability rating tasks were used but online studies have been and are being run in the context of comparing textual data to behavioral data. From the perspective taken in this paper, the main difference between off-line and online studies lies in the predictors they are trying to test: while off-line studies concentrate on the predictions of the entire model, online studies tend to – for various reasons – focus on a few variables. Bresnan and Ford (2010) use a continuous lexical decision task to test the effect of the corpus predictors context, verb, and theme in online processing across American and Australian varieties of English, while Ford and Bresnan (2013b) implement a self-paced reading task; sentences were contextualized corpus sentences. Divjak et al. (2016a) describe a self-paced reading task on attested corpus sentences to ascertain whether TAM marking, which comes out of the corpus model as the strongest predictor, plays a role in the online processing of the three most frequently used "try" verbs. Arppe et al. (2012) use eye tracking to establish the relation between the contextual probability of a *be*, *get*, or *become* passive and processing ease using authentic corpus sentences. The technicalities associated with online studies underscore the need for linguistics as a field to embrace interdisciplinary collaborations.

In general, the results of the studies reported in this paper show that an adequately constructed probabilistic model based on richly annotated corpus data can perform at a more or less equal level to human beings. The finding that the "goodness" of a corpus-based statistical model is comparable to human beings supports the claim that corpus-based models allow for a cognitively realistic language description. Neither language users nor statistical models are able to make predictions at 100 % accuracy – "language is never, ever, ever, random" (Kilgariff 2005), but it is also rarely, if ever, fully predictable

(Divjak et al. 2016b). At the same time, we need to keep in mind that models assume some kind of an ideal state of affairs, something which the real world of language never is. After all, "all models are wrong" (Box 1976: 792), some models are better than others, and the correct model can never be known with certainty (Crawley 2007: 339).

## References

Abdulrahim, Dana. 2013. *A corpus study of basic motion events in Modern Standard Arabic*. Edmonton: University of Alberta dissertation. http://hdl.handle.net/10402/era.33921 (accessed 20 January 2015)

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland & Franklin Chang. 2012. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language* 88(1). 45–81.

Antić, Eugenia. 2012. Relative frequency effects in Russian morphology. In Stefan Th. Gries & Dagmar Divjak (eds.), *Frequency effects in language learning and processing*, Vol. 1, 83–102. Berlin: De Gruyter Mouton.

Arppe, Antti. 2008. *Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy*. Helsinki: University of Helsinki dissertation. https://helda.helsinki.fi/handle/10138/19274 (accessed 28 May 2015)

Arppe, Antti. 2013a. Polytomous: Polytomous logistic regression for fixed and mixed effects. R package version 0.1.6. http://CRAN.R-project.org/package = polytomous

Arppe, Antti. 2013b. Extracting exemplars and prototypes. R vignette to accompany Divjak & Arppe (2013). http://cran.r-project.org/web/packages/polytomous/vignettes/exemplars2prototypes.pdf

Arppe, Antti & Dana Abdulrahim. 2013. Converging linguistic evidence on two flavors of production: The synonymy of Arabic COME verbs. Paper presented at Second Workshop on Arabic Corpus Linguistics, University of Lancaster, 22–26 July.

Arppe, Antti, Patrick Bolger & Dagmara Dowbor. 2012. The more evidential diversity, the merrier – contrasting linguistic data on frequency, selection, acceptability and processing. Paper presented at New Ways of Analyzing Syntactic Variation, Radboud University, Nijmegen, the Netherlands, 15–17 November.

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11(2). 295–328.

Baayen, R. Harald & Antti Arppe. 2011. Statistical classification and principles of human learning. *QITL-4-Proceedings of Quantitative Investigations in Theoretical Linguistics 4 (QITL-4)*. Berlin: Humboldt-Universität zu Berlin. http://edoc.hu-berlin.de/conferences/qitl-4/baayen-r-harald-8/PDF/baayen.pdf (accessed on 06 January 2015).

Baayen, R. Harald, Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59. 390–412.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nesset. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37. 253–291.

Barth, Danielle & Vsevolod Kapatsinski. in press. A multimodel inference approach to categorical variant choice: Construction, priming and frequency effects on the choice between full and contracted forms of *am, are* and *is*. *Corpus Linguistics and Linguistic Theory*. http://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2014-0022/cllt-2014-0022.xml (accessed 28 May 2015)

Bermel, Neil & Luděk Knittl. 2012a. Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8(2). 241–275.

Bermel, Neil & Luděk Knittl. 2012b. Morphosyntactic variation and syntactic constructions in Czech nominal declension: corpus frequency and native-speaker judgments. *Russian Linguistics* 36(1). 91–119.

Box, George E. P. 1976. Science and statistics. *Journal of the American Statistical Association* 71(356). 791–799.

Bradshaw, John. 1984. A guide to norms, ratings, and lists. *Memory & Cognition* 12(2). 202–206.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 77–96. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.) *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 186–213.

Burnham, Kenneth P. & David R. Anderson. 2002. *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd edn. New York: Springer.

Bybee, Joan L. & David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82(2). 323–355.

Caines, Andrew. 2012. 'You talking to me?' Testing corpus data with a shadowing experiment. In Stefan Th. Gries & Dagmar Divjak (eds.), *Frequency effects in language learning and processing*, 177–206. Berlin: MDe Gruyter Mouton.

Chafe, Wallace. 1992. The importance of corpus linguistics to understanding the nature of language. In Jan Svartvik (ed.), *Directions in corpus linguistics*, 79–97. Berlin: Mouton de Gruyter.

Crawley, Michael J. 2007. *The R book*. Chichester: John Wiley & Sons.

De Sutter, Gert, Dirk Speelman & Dirk Geeraerts. 2008. Prosodic and syntactic-pragmatic mechanisms of grammatical variation: The impact of a postverbal constituent on the word order in Dutch clause final verb clusters. *International Journal of Corpus Linguistics* 13(2). 194–224.

Deignan, Alice H. 2005. *Metaphor and corpus linguistics*. Amsterdam: John Benjamins.

Divjak, Dagmar. 2003. On trying in Russian: A tentative network model for near(er) synonyms. *Slavica Gandensia* 30. 25–58.

Divjak, Dagmar. 2004. *Degrees of verb integration: Conceptualizing and categorizing events in Russian*. Leuven: University of Leuven (KU Leuven) dissertation.

Divjak, Dagmar. 2010. *Structuring the lexicon: A clustered model for near-synonymy* (Cognitive Linguistics Research). Berlin: De Gruyter Mouton.

Divjak, Dagmar. 2012. Introduction. In Dagmar Divjak & Stephan Th. Gries (eds.), *Frequency effects in language*. Vol. 2: *Frequency effects in language representation*. Berlin: De Gruyter Mouton, 1–10.

Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274.

Divjak, Dagmar, Antti Arppe & Harald Baayen. 2016a. Does real language fit a self-paced reading paradigm? In Anja Gattnar, Tanja Anstatt & Christina Clasmeier (eds.), *Slavic languages in psycholinguistics*, 52–82. Tübingen: Narr Francke Attempto Verlag.

Divjak, Dagmar, Antti Arppe & Ewa Dąbrowska. 2016b. Machine meets man: Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1–33.

Divjak, Dagmar & Stefan Th. Gries. 2006. Ways of trying in Russian. Clustering behavioral profiles. *Journal of Corpus Linguistics and Linguistic Theory* 2(1). 23–60.

Divjak, Dagmar & Stefan Th. Gries. 2008. Clusters in the mind? Converging evidence from near-synonymy in Russian. *The Mental Lexicon* 3(2). 188–213.

Divjak, Dagmar & Stefan Th. Gries (eds.). 2012. *Frequency effects in language*. Vol. 2: *Frequency effects in language representation*. Berlin: De Gruyter Mouton.

Erker, Daniel & Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88(3). 526–557.

Ford, Marilyn & Joan Bresnan. 2013a. Using convergent evidence from psycholinguistics and usage. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 295–312. Cambridge: Cambridge University Press.

Ford, Marilyn & Joan Bresnan. 2013b. 'They whispered me the answer' in Australia and the US: A comparative experimental study. In Tracy Holloway King & Valeria de Paiva (eds.), *From quirky case to representing space: Papers in honor of Annie Zaenen*, 95–107. Stanford: CSLI Publications. http://web.stanford.edu/group/cslipublications/cslipublications/Online/azfest-final.pdf (accessed 22 January 2015).

Frary, Robert B. 1988. Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice* 7(2). 33–38.

Gilquin, Gaëtanelle & Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1). 1–26.

Glynn, Dylan & Kerstin Fischer (eds.). 2010. *Quantitative methods in cognitive semantics: Corpus-driven approaches* (Cognitive Linguistics Research 46). Berlin: De Gruyter Mouton.

Glynn, Dylan & Justyna Robinson (eds.). 2014. *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (Human Cognitive Processing 43). Amsterdam: John Benjamins.

Gries, Stefan Th. 2003. *Multifactorial analysis in corpus linguistics: A study of particle placement*. London: Continuum Press.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2010. Converging evidence II: More on the association of verbs and constructions. In Sally Rice & John Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 59–72. Stanford, CA: Center for the Study of Language and Information.

Gries, Stefan Th. & Martin Hilpert. 2010. Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14(3). 293–320.

Gries, Stefan Th. & Dagmar Divjak (eds.). 2012. *Frequency effects in language*. Vol. 1: *Frequency effects in language learning and processing*. Berlin: De Gruyter Mouton.

Grondelaers, Stefan & Dirk Speelman. 2007. A variationist account of constituent ordering in presentative sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory* 3(2). 161–193.

Harrell, Frank E. 2001. *Regression modeling strategies: With applications to linear models, logistic regression and survival analysis*. New York: Springer.

Hosmer, David W., Jr., Stanley Lemeshow & Rodney X. Sturdivant. 2013. *Applied logistic regression*. Hoboken, NJ: John Wiley & Sons.

Jaeger, T. Florian 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language* 59(4). 434–446.

Jurafsky, Dan. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 39–95. Cambridge, MA: MIT Press.

Kendall, Tyler, Joan Bresnan & Gerard Van Herk. 2011. The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory* 7(2). 229–244.

Kilgariff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276.

Klavan, Jane 2012. *Evidence in linguistics: Corpus-linguistic and experimental methods for studying grammatical synonymy*. (Dissertationes Linguisticae Universitatis Tartuensis). Tartu: University of Tartu Press.

Klavan, Jane. 2014. How good is good? Evaluating the performance of probabilistic statistical classification models for predicting constructional choices. Paper presented at 5th UK Cognitive Linguistics Conference, University of Lancaster, 29–31 July.

Kotz, Samuel (ed.). 2006. *Encyclopedia of statistical sciences*, Vol. 11. Hoboken, NJ: Wiley and Sons.

McEnery, Tony & Andrew Hardie 2012. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević & R. Harald Baayen. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4).

Mitchell, Gregory. 2012. Revisiting truth or triviality the external validity of research in the psychological laboratory. *Perspectives on Psychological Science* 7(2). 109–117.

Pinheiro, José C. & Douglas M. Bates. 2000. *Mixed-effects models in S and S-PLUS*. New York: Springer.

Raymond, William D. & Esther L. Brown. 2012. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In Stefan Th. Gries & Dagmar Divjak (eds.), *Frequency effects in language learning and processing*, 35–52. Berlin: De Gruyter Mouton.

Resnik, Philip & Jimmy Lin. 2010. Evaluation of NLP systems. In Alexander Clark, Chris Fox & Shalom Lappin (eds.), *The handbook of computational linguistics and natural language processing*, 271–295. Oxford: Wiley-Blackwell.

Roland, Douglas, Jeffrey L. Elman & Victor S. Ferreira. 2006. Why is *that*? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition* 98. 245–272.

Sankoff, David & William Labov. 1979. On the uses of variable rules. *Language in Society* 8(3). 189–222.

Szmrecsanyi, Benedikt. 2013. Diachronic probabilistic grammar. *English Language and Linguistics* 19(3). 41–68.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen & Hans van Halteren. 2013. Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2). 227–262.

Tooley, Kristen M. & Kathryn Bock. 2014. On the parity of structural persistence in language production and comprehension. *Cognition* 132(2). 101–136.

Van de Weijer, Joost, Carita Paradis, Caroline Willners & Magnus Lindgren. 2012. As lexical as it gets: The role of co-occurrence of antonyms in a visual lexical decision experiment. In Dagmar Divjak & Stefan Th. Gries (eds.), *Frequency effects in language representation*, 255–279. Berlin: De Gruyter Mouton.

Wasow, Thomas & Jennifer Arnold. 2003. Post-verbal constituent ordering in English. *Topics in English Linguistics* 43. 119–154.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.